UNIVERSIDAD PRIVADA SAN JUAN BAUTISTA

FACULTAD DE CIENCIAS DE LA SALUD

ESCUELA PROFESIONAL DE MEDICINA HUMANA

PC4

ASIGNATURA:

SISTEMATIZACIÓN Y MÉTODOS ESTADÍSTICOS

DOCENTE:

SEGUNDO VICENTE CASTRO LOPEZ

ESTUDIANTE:

Arianna Samantha Caballero Martinez

Lozano Laura Marina

Lizbeth Adriana Pachas Rojas

Nayely Luz Rojas Cortez

Alexander Manay Ventura

Perú

2025

GRUPO 5

SEMANA 13

INTEGRANTES:

- Arianna Samantha Caballero Martinez
- Lozano Laura Marina
- Lizbeth Adriana Pachas Rojas
- Nayely Luz Rojas Cortez
- Alexander Manay Ventura

Cargar los paquetes

```{r}
install.packages("factoextra")
install.packages("cluster")
```

```{r}
library(factoextra)
library(cluster)
library(here)
library(rio)
library(tidyverse)
```

1 ¿Cómo aplicaremos Machine Learning a esta sesión?

Para intentar responder preguntas de investigación en el contexto de enfermedades hepáticas como la cirrosis, es necesario contar con múltiples mediciones clínicas en una misma cohorte de pacientes. En este caso, además de registrar variables habituales como la edad, sexo y el estado clínico final (fallecimiento o censura), se han recolectado numerosos parámetros bioquímicos y clínicos: bilirrubina, colesterol, albúmina, cobre sérico, triglicéridos, plaquetas, tiempo de protrombina, entre otros.

La complejidad de este tipo de datos reside en que muchas de estas variables pueden estar interrelacionadas. Por ejemplo, pacientes con etapas avanzadas de cirrosis tienden a presentar niveles alterados en múltiples marcadores hepáticos y hematológicos, lo que genera una dependencia entre variables. Si analizáramos estas variables por separado, podríamos perder patrones importantes presentes en el conjunto de datos multivariado.

Una forma común de abordar este problema en estadística tradicional es excluir variables correlacionadas o con "poca variabilidad", pero esta estrategia puede llevar a una pérdida significativa de información relevante. Por ello, en esta sesión aplicaremos técnicas de machine learning no supervisado, como el análisis de componentes principales (PCA) y el agrupamiento (clustering), que nos permitirán reducir la dimensionalidad del dataset y, al mismo tiempo, identificar grupos de pacientes que comparten características clínicas similares.

1.1 Uso de las técnicas de agrupamiento para responden preguntas de investigación en salud

Las técnicas de agrupamiento son un tipo de técnica exploratoria que puede usarse con el objetivo de clasificar observaciones (por ejemplo pacientes que forman parte de una muestra) en grupos en base a su similaridad y desimilaridad de las variables. A partir de esto, obtendremos grupos cuyos individuos que pertenecen a un mismo grupo son similares pero diferentes a individuos que pertenecen a otros grupos.

Los grupos encontrados pueden ser usados para hacer predicciones o evaluar diferencias en parámetros de laboratorio. Por ejemplo, entre grupos encontrados de pacientes quienes iniciaron su tratamiento para el cáncer, podemos comparar su supervivencia, calidad de vida luego de dos años u otras medidas a partir de los clusters (grupos) encontrados.

2 Análisis de agrupamiento herarquico (Hierarchical Clustering)

2.1 Sobre el problema para esta sesión

El dataset de esta sesión contiene información clínica y de laboratorio de 418 pacientes con diagnóstico de cirrosis hepática. Incluye variables como bilirrubina, albúmina, cobre sérico, colesterol, fosfatasa alcalina, plaquetas, entre otras, así como características clínicas como ascitis, edema y presencia de aracnoides. El objetivo de este ejercicio es aplicar técnicas de aprendizaje no supervisado, como análisis de componentes principales (PCA) y agrupamiento K-means, para identificar grupos de pacientes con perfiles clínicos similares, lo cual podría facilitar la identificación de patrones de evolución o categorías de riesgo diferenciadas.

2.2 El dataset para esta sesión

Para ilustrar el proceso de análisis usaremos un dataset que contiene información de 418 pacientes con diagnóstico de cirrosis hepática. El conjunto de datos incluye variables clínicas, demográficas y de laboratorio, tales como: edad (en días), sexo (hombre/mujer), duración del seguimiento (en días), estado al final del seguimiento (fallecido, censurado, censurado por trasplante) y tipo de tratamiento recibido (D-penicilamina o placebo).

Además, se consideran variables clínicas como presencia de ascitis, hepatomegalia, aracnoides y edema (todas categóricas), y parámetros bioquímicos como: bilirrubina (mg/dL), colesterol (mg/dL), albúmina (g/dL), cobre (µg/dL), fosfatasa alcalina (U/L), SGOT (U/L), triglicéridos (mg/dL), recuento de plaquetas (miles/µL) y tiempo de protrombina (segundos). Finalmente, se incluye la clasificación clínica de la etapa de la enfermedad (Etapa 1 a Etapa 4), codificada como variable categórica ordinal.

Este conjunto de variables permitirá realizar un análisis multivariado con reducción de dimensionalidad y técnicas de agrupamiento para explorar patrones clínicos en esta población.

2.2.1 Importando los datos

```{r}
cirrosis_data <- import(here("data", "cirrosis.csv"))
```

```
Registered S3 method overwritten by 'data.table':
  method           from
  print.data.table
```

Data	
cirrosis_data	418 obs. of 20 variables

2.3 Preparación de los datos

2.3.1 Solo datos numéricos

Para el análisis de agrupamiento de esta sesión usaremos únicamente las variables numéricas del dataset. Aunque es posible incluir variables categóricas mediante codificación, en este caso nos enfocaremos solo en variables numéricas. Por ello, se eliminarán columnas categóricas como `Sexo`, `Estado`, `Medicamento`, `Ascitis`, `Hepatomegalia`, `Aracnoides`, `Edema` y `Etapa`. La columna `ID` se utilizará como identificador.

```{r}
cirrosis_data_1 <- cirrosis_data |>
  select(-Sexo, -Estado, -Medicamento, -Ascitis, -Hepatomegalia,
         -Aracnoides, -Edema, -Etapa) |>
  column_to_rownames("ID")
```

Data	
cirrosis_data	418 obs. of 20 variables
cirrosis_dat…	418 obs. of 11 variables

2.3.2 La importancia de estandarizar

Antes de aplicar técnicas de agrupamiento, es esencial estandarizar las variables numéricas del dataset, ya que muchas de ellas están medidas en distintas escalas (por ejemplo, bilirrubina en mg/dL, cobre en µg/dL, tiempo de protrombina en segundos).

Sin una estandarización previa, las variables con valores numéricos más grandes pueden dominar el cálculo de distancias y sesgar la formación de grupos. Por eso, aplicaremos la función `scale()` en R, que transforma las variables para que tengan media cero y desviación estándar uno, asegurando que todas contribuyan de forma equitativa al análisis.

```r
{r}
cirrosis_data_escalado = scale(cirrosis_data_1)
```

Un vistazo a los datos antes del escalamiento:

```r
{r}
head(cirrosis_data_1)
```

Description: df [6 × 11]

	Dias_Seg... <int>	Edad <int>	Bilirrubina <dbl>	Colesterol <int>	Albumina <dbl>	Cobre <int>
1	400	21464	14.5	261	2.60	156
2	4500	20617	1.1	302	4.14	54
3	1012	25594	1.4	176	3.48	210
4	1925	19994	1.8	244	2.54	64
5	1504	13918	3.4	279	3.53	143
6	2503	24201	0.8	248	3.98	50

6 rows | 1-7 of 11 columns

y un vistazo después del escalamiento:

```r
{r}
head(cirrosis_data_escalado)
```

```
  Dias_Seguimiento       Edad Bilirrubina Colesterol    Albumina      Cobre
Fosfatasa_Alcalina       SGOT Trigliceridos  Plaquetas Tiempo_Protrombina
1     -1.373965243  0.7680208  2.55908571 -0.4678298 -2.11176507  0.6815669
-0.1236485  0.2714953     0.7259994 -0.6816595          1.4366622
2      2.337540359  0.5460516 -0.48118215 -0.2910634  1.51200645 -0.5098282
2.5285799 -0.1593725    -0.5633599 -0.3663804          -0.1288950
3     -0.819955139  1.8503499 -0.41311645 -0.8342967 -0.04103849  1.3123054
-0.6852287 -0.4666061    -1.0698938 -1.0783009          1.2409675
4      0.006533792  0.3827850 -0.32236219 -0.5411232 -2.25295097 -0.3930247
1.9338282 -1.0921846    -0.5019618 -0.7528516          -0.4224370
5     -0.374574466 -1.2095228  0.04065487 -0.3902250  0.07661643  0.5297224
-0.6128119 -0.1658981    -0.8089521 -1.2308553          0.1646469
6      0.529765557  1.4852931 -0.54924785 -0.5238777  1.13551071 -0.5565495
-0.4852650 -0.5212803    -0.9470977          NA          0.2624943
```

INTERPRETACIÓN DE LOS RESULTADOS

Cada celda representa el valor estandarizado de una variable para un paciente específico. Un valor positivo indica que el paciente presenta un resultado por encima del promedio poblacional, mientras que un valor negativo indica un resultado por debajo del promedio.

La primera fila:

El paciente presenta un valor considerablemente alto de bilirrubina (+2.55), lo cual podría reflejar una alteración hepática significativa.
Su colesterol (−0.47) está por debajo del promedio del grupo.
El tiempo de protrombina (+1.43) también está aumentado, lo cual podría estar asociado a una disfunción hepática avanzada.
Este tipo de estandarización permite que variables medidas en distintas unidades (mg/dL, µg/dL, segundos, etc.) sean comparables en escala, evitando que aquellas con mayor rango numérico dominen los análisis posteriores.

2.4 Cálculo de distancias

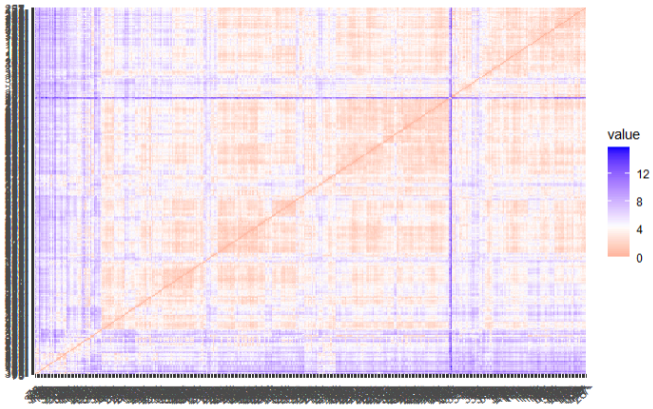
Dado que uno de los pasos es encontrar "cosas similares", necesitamos definir "similar" en términos de distancia. Esta distancia la calcularemos para cada par posible de objetos (participantes) en nuestro dataset. Por ejemplo, si tuvieramos a los pacientes A, B y C, las distancia se calcularían para A vs B; A vs C; y B vs C. En R, podemos utilizar la función `dist()` para calcular la distancia entre cada par de objetos en un conjunto de datos. El resultado de este cálculo se conoce como matriz de distancias o de disimilitud.

```
{r}
dist_cirrosis_data <- dist(cirrosis_data_escalado, method = "euclidean")
```

2.4.1 Visualizando las distancias euclidianas con un mapa de calor

Una forma de visualizar si existen patrones de agrupamiento es usando mapas de calor (heatmaps). En R usamos la función `fviz_dist()` del paquete factoextra para crear un mapa de calor.

```
{r}
fviz_dist(dist_cirrosis_data)
```



INTERPRETACIÓN DE LOS RESULTADOS

Se observan zonas definidas con menor distancia rosadas claras lo que sugiere que existen grupos de pacientes con características clínicas similares.

Las franjas azules verticales y horizontales indican pacientes que son notablemente distintos del resto, posiblemente casos atípicos o con perfiles clínicos extremos.

La diagonal principal siempre marca distancia cero

2.5 El método de agrupamiento: función de enlace (linkage)

El agrupamiento jerárquico es una técnica que permite organizar a los pacientes según la similitud de sus perfiles clínicos, comenzando por agrupar aquellos que presentan características más parecidas entre sí. Este método parte de la matriz de distancias previamente calculada a partir de las variables clínicas estandarizadas.

Una vez agrupados los pacientes más similares, se forman nuevos grupos (clústeres), y es necesario decidir cómo calcular la distancia entre estos nuevos grupos y el resto de los individuos o clústeres ya existentes. Para ello, se emplea una función de enlace (*linkage*), la cual define el criterio para unir clústeres durante el proceso jerárquico.

Existen distintas funciones de enlace, como el enlace simple, enlace completo, enlace promedio, centroide, y el método de varianza mínima de Ward, entre otros. En este análisis, se utiliza el método de Ward, que tiene la ventaja de minimizar la varianza dentro de cada clúster y tiende a generar grupos de tamaño equilibrado, facilitando así la interpretación clínica de los resultados.

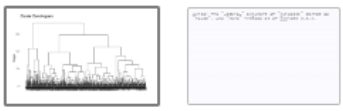
Este proceso se repite hasta que todos los pacientes quedan agrupados en un único árbol jerárquico o dendrograma, el cual puede ser visualizado para decidir cuántos clústeres resultan clínicamente relevantes.

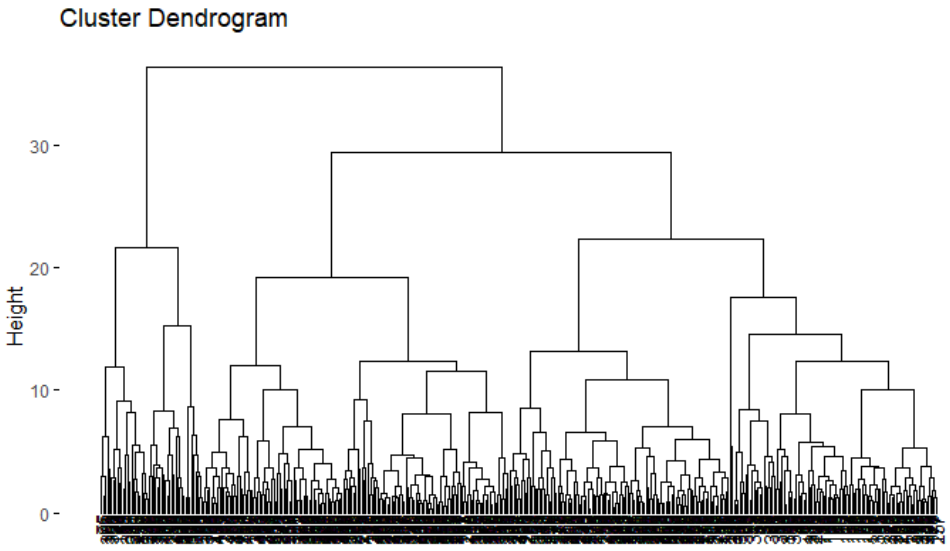
```{r}
dist_link_cirrosis_data <- hclust(d = dist_cirrosis_data, method = "ward.D2")
```

2.7 Dendrogramas para la visualización de patrones

Los dendrogramas es una representación gráfica del árbol jerárquico generado por la función `hclust()`.

```{r}
fviz_dend(dist_link_cirrosis_data, cex = 0.7)
```





INTERPRETACIÓN DE LOS RESULTADOS

Las ramas del dendrograma muestran cómo se agrupan progresivamente los pacientes en función de su similitud.

A medida que se asciende en el gráfico, se fusionan clústeres más grandes, y el eje vertical (Height) indica la distancia (disimilitud) entre grupos.

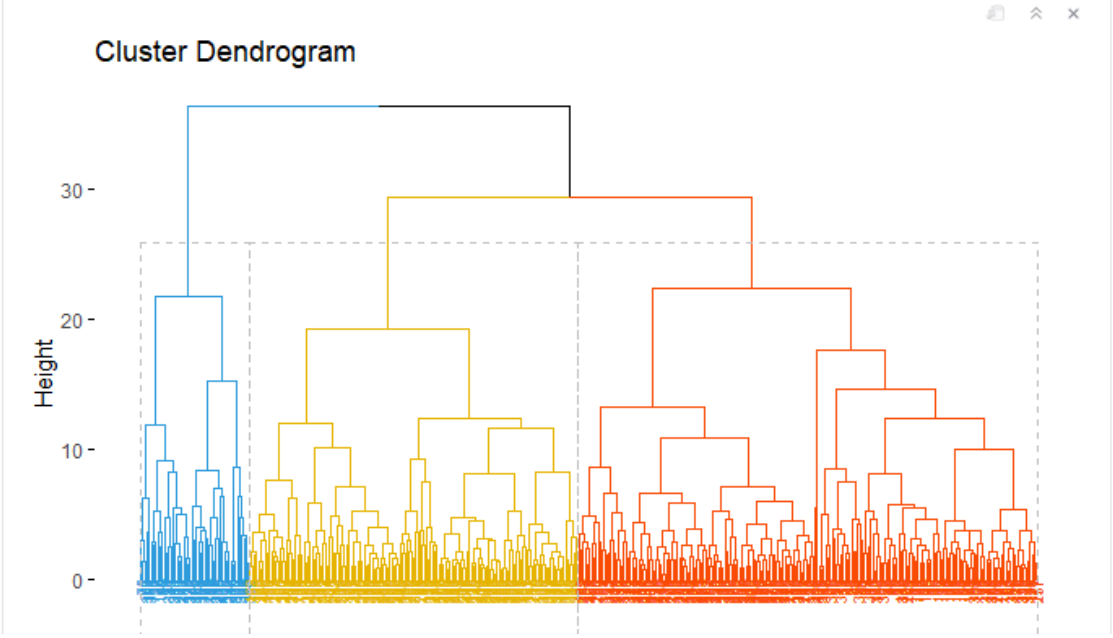
Se pueden observar algunos cortes naturales en la estructura, lo que sugiere que existen grupos distinguibles dentro de la población analizada.

2.8 ¿Cúantos grupos se formaron en el dendrograma?

Uno de los retos del agrupamiento jerárquico es que el método no determina automáticamente cuántos grupos (clústeres) se deben formar. La decisión de dónde "cortar" el dendrograma depende del criterio del investigador, considerando la estructura del gráfico y el propósito clínico del análisis.

En el caso del dendrograma obtenido para los pacientes con cirrosis, se observan varias separaciones naturales. Para fines exploratorios, se ha decidido cortar el árbol en **tres grupos**, lo cual permite identificar subpoblaciones con perfiles clínicos diferenciados.

```r
{r}
fviz_dend(dist_link_cirrosis_data,
          k = 3,
          cex = 0.5,
          k_colors = c("#2E9FDF", "#E7B800", "#FC4E07"),
          color_labels_by_k = TRUE,
          rect = TRUE)
```



INTERPRETACIÓN DE LOS RESULTADOS
Cada color (azul, amarillo y rojo) representa un grupo de pacientes con perfiles clínicos similares entre sí.

Las diferencias de altura entre los grupos sugieren que hay una separación significativa entre estos clústeres, lo que respalda la existencia de subpoblaciones diferenciadas dentro de la cohorte de pacientes con cirrosis.

El grupo azul, por ejemplo, muestra ramas más cortas y compactas, lo cual podría indicar pacientes con perfiles homogéneos. En contraste, el grupo rojo presenta mayor dispersión interna, lo que sugiere mayor heterogeneidad clínica.

3 Agrupamiento con el algoritmo K-Means

El método de agrupamiento (usando el algoritmo) K-means es la técnica de machine learning más utilizado para dividir un conjunto de datos en un número determinado de k grupos (es decir, k clústeres), donde k representa el número de grupos predefinido por el investigador. Esto contrasta con la técnica anterior, dado que aquí sí iniciamos con un grupo pre-definido cuya idoniedad (de los grupos) puede ser evaluado. En detalle, el esta técnica clasifica a los objetos (participantes) del dataset en múltiples grupos, de manera que los objetos dentro de un mismo clúster sean lo más similares posible entre sí (alta similitud intragrupo), mientras que los objetos de diferentes clústeres sean lo más diferentes posible entre ellos (baja similitud intergrupo). En el agrupamiento k-means, cada clúster se representa por su centro (centroide), que corresponde al promedio de los puntos asignados a dicho clúster.

Aquí como funciona el algoritmo de K-Means

1. Indicar cuántos grupos (clústeres) se quieren formar. Por ejemplo, si se desea dividir a los pacientes en 3 grupos según sus características clínicas, entonces K=3.
2. Elegir aleatoriamente K casos del conjunto de datos como centros iniciales. Por ejemplo, R selecciona al azar 3 pacientes cuyas características (edad, IMC, creatinina, etc.) servirán como punto de partida para definir los grupos.
3. Asignar cada paciente al grupo cuyo centro esté más cerca, usando la distancia euclidiana. Es como medir con una regla cuál centroide (paciente promedio) está más próximo a cada paciente en función de todas sus variables.
4. Calcular un nuevo centro para cada grupo. Es decir, calcular el promedio de todas las variables de los pacientes que quedaron en ese grupo. Por ejemplo, si en el grupo 1 quedaron 40 pacientes, el nuevo centroide será el promedio de la edad, IMC, creatinina, etc., de esos 40 pacientes. Este centroide es un conjunto de valores (uno por cada variable).
5. Repetir los pasos 3 y 4 hasta que los pacientes dejen de cambiar de grupo o hasta alcanzar un número máximo de repeticiones (en R, por defecto son 10 repeticiones). Esto permitirá que los grupos finales sean estables.

3.1 El problema y dataset para este ejercicio

Usaremos el mismo dataset y el mismo problema que el que empleamos en el ejercicio anterior (para Agrupamiento Jerárquico).

3.2 Estimando el número óptimo de clusters

Como indiqué arriba, el método de agrupamiento k-means requiere que el usuario especifique el número de clústeres (grupos) a generar. Una pregunta fundamental es: ¿cómo elegir el número adecuado de clústeres esperados (k)?

Aquí muestro una solución sencilla y popular: realizar el agrupamiento k-means probando diferentes valores de k (número de clústeres). Luego, se grafica la suma de cuadrados dentro de los clústeres (WSS) en función del número de clústeres. En R, podemos usar la función fviz_nbclust() para estimar el número óptimo de clústeres.

Primero nos aseguramos que las variables sean numéricas:

```r
cirrosis_data_num <- cirrosis_data_1 |> dplyr::select(where(is.numeric))
```

Eliminamos filas con NA SOLO después de seleccionar variables numéricas:

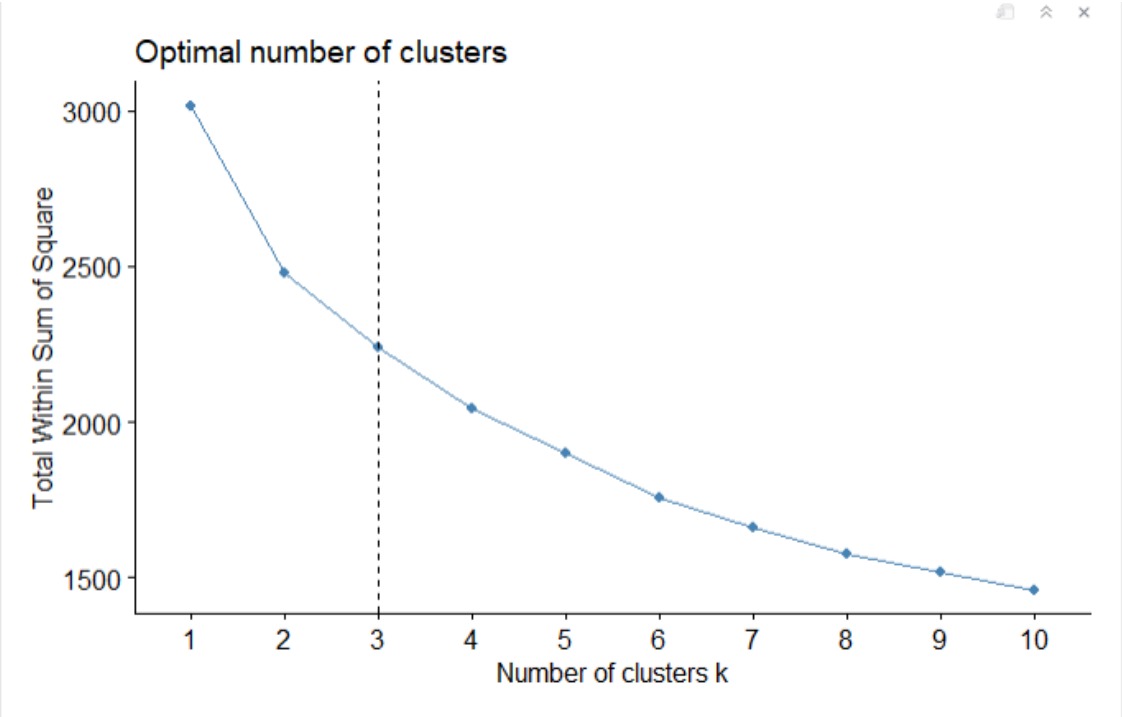
```r
cirrosis_data_limpio <- na.omit(cirrosis_data_num)
```

Ahora podemos escalar los datos:

```r
cirrosis_data_escalado <- scale(cirrosis_data_limpio)
```

Ahora graficamos la suma de cuadrados dentro de los gráficos

```r
fviz_nbclust(cirrosis_data_escalado, kmeans, nstart = 25, method = "wss") +
  geom_vline(xintercept = 3, linetype = 2)
```



INTERPRETACIÓN DE LOS RESULTADOS

El eje vertical muestra la suma total de las distancias al centroide dentro de cada clúster, mientras que el eje horizontal representa el número de clústeres.

A medida que aumenta el número de clústeres, la WSS disminuye, ya que los grupos se ajustan mejor a los datos.
Sin embargo, a partir de k = 3 la reducción de la WSS comienza a desacelerarse notablemente, formando un punto de inflexión (codo) claro.
Este punto de codo sugiere que tres clústeres representan una solución óptima de compromiso entre simplicidad y capacidad de agrupamiento, lo que respalda la elección visual ya realizada en el dendrograma jerárquico.

3.3 Cálculo del agrupamiento k-means

Dado que el resultado final del agrupamiento k-means es sensible a las asignaciones aleatorias iniciales, se especifica el argumento `nstart = 25`. Esto significa que R intentará 25 asignaciones aleatorias diferentes y seleccionará la mejor solución, es decir, aquella con la menor variación dentro de los clústeres. El valor predeterminado de `nstart` en R es 1. Sin embargo, se recomienda ampliamente utilizar un valor alto, como 25 o 50, para obtener un resultado más estable y confiable. El valor empleado aquí, fue usado para determinar el número de clústeres óptimos.

```{r}
set.seed(123)
km_res <- kmeans(cirrosis_data_escalado, 3, nstart = 25)
```

```{r}
km_res
```

```
K-means clustering with 3 clusters of sizes 169, 70, 37

Cluster means:
  Dias_Seguimiento       Edad Bilirrubina Colesterol   Albumina      Cobre
Fosfatasa_Alcalina        SGOT Trigliceridos  Plaquetas Tiempo_Protrombina
1        0.5087735 -0.4006207  -0.4116397 -0.2225916  0.4118981 -0.3856295
-0.10815495 -0.21275785    -0.1927009  0.1935875         -0.3257476
2       -0.6491871  0.8021379   0.1032105 -0.2608838 -0.6387255  0.4774946
-0.01703632  0.05264236    -0.1695352 -0.4223170          0.5797992
3       -0.6811590 -0.3602193   1.8759759  1.5665568 -0.3329499  1.1298395
0.57487328  1.07780835     1.2325334  0.2748978          0.4185211
```

```
0.57487328  1.07780835     1.2325334  0.2748978          0.4185211

Clustering vector:
  1   2   3   4   5   7   8   9  10  11  12  13  15  16  17  18  19  20  21  22
23  24  25  26  27  28  29  30  31  32  33  34  35  36  37  38  39  43  44  46  47
48  50
  2   1   2   2   1   1   1   1   2   1   2   1   1   1   2   3   1   2   1   2
3   1   1   3   3   3   1   2   1   1   2   1   1   1   2   1   1   1   1   3   1
1   1
 51  52  54  55  56  57  59  60  61  62  63  64  65  66  67  68  69  71  72  73
74  75  76  77  78  79  80  81  82  83  84  85  86  87  88  89  90  91  92  93  94
97  98
  1   3   2   2   3   2   1   1   1   2   3   2   1   1   1   1   3   1   1   1
3   3   3   2   2   1   2   3   1   1   1   1   3   1   1   1   1   2   2   1   2
2   1
 99 100 101 102 103 104 105 107 108 109 110 111 112 113 114 115 116 117 118 119
120 121 122 124 125 127 130 131 132 133 134 135 136 137 138 139 140 141 142 143
144 145 147
  1   2   1   1   2   1   1   2   1   1   1   1   1   2   1   1   1   3   1   1
3   2   1   1   1   1   3   2   1   1   1   1   1   1   3   2   1   1   1   1   3
1   2
148 149 151 152 153 154 155 156 157 158 159 160 161 162 163 165 166 167 169 170
172 173 175 177 179 180 181 183 184 185 186 187 188 189 191 192 193 194 195 196
197 198 199
  3   2   1   2   1   2   1   3   1   1   2   1   1   2   1   2   3   2   1   1
1   1   1   1   1   1   1   1   3   1   2   3   2   1   3   1   3   1   1   2   1
1   1
200 201 202 203 204 206 208 209 210 212 213 214 215 217 219 220 221 222 223 224
225 226 227 228 229 230 231 232 233 234 235 236 237 239 240 241 242 243 244 245
246 247 248
  1   1   1   1   1   1   2   1   1   1   1   2   3   3   1   2   1   2   2   1
1   1   2   1   2   1   2   1   2   1   3   1   2   2   1   3   2   2   3   1   1
3   1
249 250 251 252 253 254 255 256 257 258 259 260 262 263 264 265 266 267 268 269
270 271 272 273 275 276 277 278 279 280 281 282 283 284 285 286 287 288 289 290
291 292 293
  1   1   1   2   2   1   1   1   1   1   2   2   1   1   2   1   1   2   2   2
1   1   1   1   1   1   1   1   1   1   3   1   1   1   1   1   2   1   2   1   1
2   2
294 295 296 297 298 299 301 302 303 304 305 306 307 308 309 310 311 312
  3   1   1   2   1   1   1   1   2   1   3   1   1   1   2   2   1   3

Within cluster sum of squares by cluster:
[1] 1008.8319  554.5053  677.5751
 (between_SS / total_SS =  25.7 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
"betweenss"    "size"         "iter"         "ifault"
```

INTERPRETACIÓN DE LOS RESULTADOS

Se aplicó el algoritmo de K-means con *k = 3* para agrupar a los pacientes con cirrosis según sus características clínicas estandarizadas. El análisis identificó tres grupos con tamaños de 169, 70 y 37 pacientes, respectivamente. El Clúster 1 presenta valores cercanos al promedio general, posiblemente representando pacientes con perfil clínico intermedio. El Clúster 2 muestra reducciones en variables como albúmina, plaquetas y tiempo de protrombina, lo cual podría reflejar un estado clínico más comprometido. En contraste, el Clúster 3 se caracteriza por elevaciones marcadas en bilirrubina, colesterol, cobre y enzimas hepáticas, lo que sugiere un subgrupo con mayor alteración hepática o fase avanzada de la enfermedad.

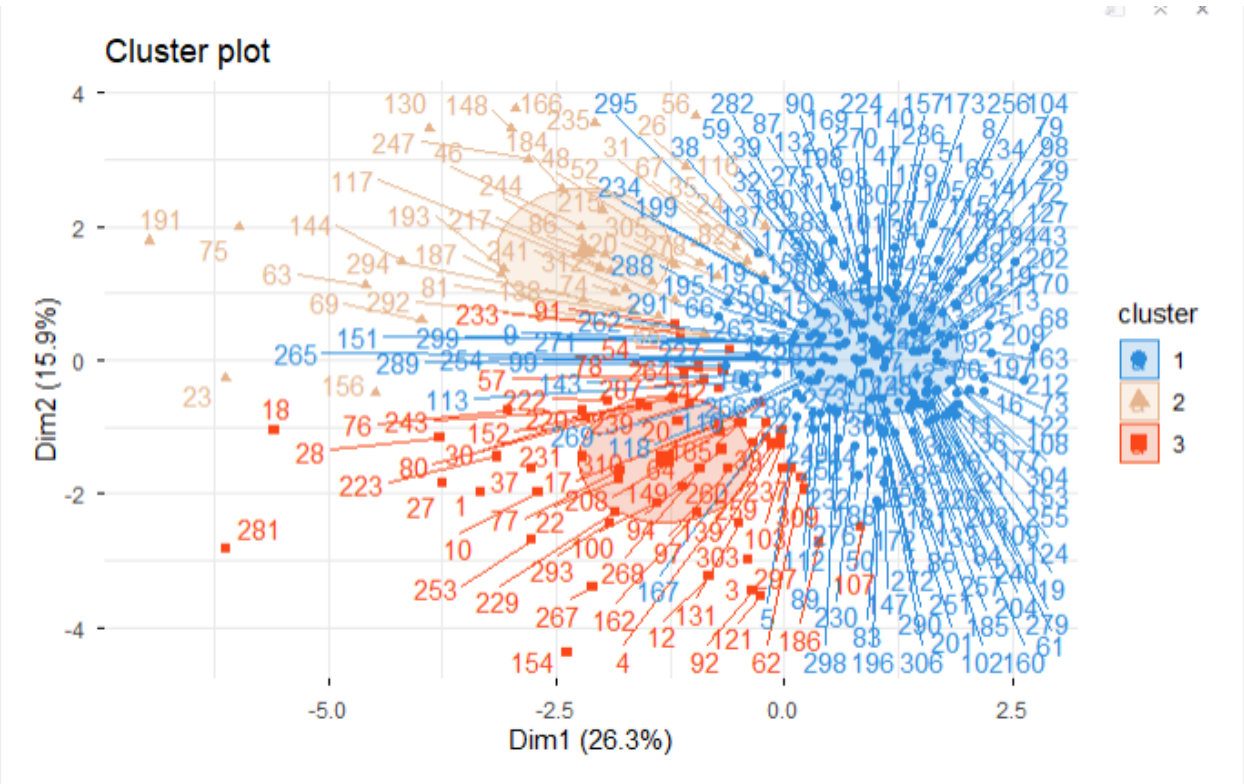
Este agrupamiento constituye una base útil para describir perfiles clínicos diferenciados dentro de la cohorte y explorar asociaciones con desenlaces relevantes.

3.4 Visualización de los clústeres k-means

Al igual que el análisis anterior, los datos se pueden representar en un gráfico de dispersión, coloreando cada observación o paciente según el clúster al que pertenece. El problema es que los datos contienen más de dos variables, y surge la pregunta de qué variables elegir para representar en los ejes X e Y del gráfico. Una solución es reducir la cantidad de dimensiones aplicando un algoritmo de reducción de dimensiones, como el Análisis de Componentes Principales (PCA). El PCA transforma las 52 variables originales en dos nuevas variables (componentes principales) que pueden usarse para construir el gráfico.

La función `fviz_cluster()` del paquete factoextra se puede usar para visualizar los clústeres generados por k-means. Esta función toma como argumentos los resultados del k-means y los datos originales (hemo_data_escalado).

```{r}
fviz_cluster(
  km_res,
  data = cirrosis_data_escalado,
  palette = c("#2E8FDF", "#E7B890", "#FC4A19"),
  ellipse.type = "euclid",
  repel = TRUE,
  ggtheme = theme_minimal()
)
```



INTERPRETACIÓN DE LOS RESULTADOS

Se observan tres grupos claramente diferenciados, confirmando visualmente que la elección de $k = 3$ fue adecuada.
El Clúster 1 (azul) agrupa a la mayoría de los pacientes, concentrados en la parte derecha del gráfico, lo que sugiere un grupo clínicamente más homogéneo.
El Clúster 2 (amarillo) se distribuye hacia la parte inferior izquierda, y representa una subpoblación con un perfil clínico distinto, posiblemente de menor severidad.
El Clúster 3 (rojo) aparece más disperso en la parte superior izquierda, indicando mayor variabilidad interna, y puede estar asociado a perfiles clínicos más complejos o descompensados.